nature
biotechnology

Exploiting biological complexity for strain improvement through systems biology

Gregory Stephanopoulos, Hal Alper & Joel Moxley

Cellular complexity makes it difficult to build a complete understanding of cellular function but also offers innumerable possibilities for modifying the cellular machinery to achieve a specific purpose. The exploitation of cellular complexity for strain improvement has been a challenging goal for applied biological research because it requires the coordinated understanding of multiple cellular processes. It is therefore pursued most efficiently in the framework of systems biology. Progress in strain improvement will depend not only on advances in technologies for high-throughput measurements but, more importantly, on the development of theoretical methods that increase the information content of these measurements and, as such, facilitate the elucidation of mechanisms and the identification of genetic targets for modification.

Although the term 'systems biology' entered the popular lexicon only recently, the concept of an integrated, systemic approach to the analysis and optimization of cellular processes has been applied routinely by engineers and scientists for many years. The expanded view of the cell, made possible by genome sequencing and parallel, high-throughput technologies for measuring the relative abundance of important classes of intracellular molecules, revealed the obvious: hundreds or thousands of molecules, previously excluded from the focus of research, were found to vary significantly in response to a simple genetic or environmental perturbation. Systems biology soon emerged as a term and a field of scientific inquiry to describe an approach that considers genome-scale and cell-wide measurements in elucidating biological processes and mechanisms.

Traditionally, cells were considered as elegant systems of immense complexity that were, nevertheless, well-coordinated and optimized for a particular purpose. Research efforts by necessity were narrowly focused, leading over the years to the rigorous understanding of various specific, low-level processes. Recently, genome sequencing and related technologies have, for several organisms, provided a window to the broad biomolecular landscape underlying cellular phenotype. In addition to being able to quantify the abundance of important classes of biological molecules, we can now probe the interactions among them. Systems biology aims to interpret and contextualize large, diverse sets of biological measurements and elucidate the mechanisms behind complex phenomena through an integrated perspective. This is a formidable task. To maximize the probability of success, we must anchor systems biology analyses to specific questions and build upon the existing core infrastructure created by targeted studies. Strain improvement represents a specific goal suitable for the application of systems biology. However, it remains to be seen whether the understanding gained will expedite forward engineering, that is, intelligent modifications based upon system understanding of cellular processes and justify the investment required for system characterization.

Cellular complexity is a manifestation of the enormous diversity of molecules and reaction processes needed to carry out cellular functions. This review focuses on the exploitation of complexity for strain improvement within a systems biology framework. We consider several questions of central importance to this goal, such as identifying genetic targets for modification to improve strains, elucidating biomolecular interaction networks, increasing the information content of large-scale metabolomic measurements, and integrating genomic/metabolic data. We also offer a vision for the future of systems biology in strain improvement and close with a word of caution regarding the expectations of this promising field, which has been overplayed on occasion without accounting for the many difficulties that still exist and need to be resolved before the potential of this field is fully realized.

An iterative framework for accumulating systems insight

A fundamental premise of systems-based research is that the underlying mechanisms and interactions of a biological system can be probed by introducing a variety of perturbations and measuring the system response. Figure 1 illustrates the accumulation of insight from a systems-biology cycle for strain improvement. We iteratively accumulate systems insight in two steps.

First, perturbations can be introduced into cellular networks and environments to create an altered phenotype. In recognition of the importance of these changes in probing the genotype-phenotype relationship, a diverse set of tools has emerged to create gene deletions and amplifications, which are used in conjunction with altered environments. Molecular biology advances have made it possible to perform these modifications at will. In addition to gene-specific techniques, several combinatorial tools have been developed, which, when combined with high-throughput screening, allow for randomized gene-expression levels (including deletions) and genomic library complementation[1,2].

Department of Chemical Engineering, Massachusetts Institute of Technology, Room 56-469, Cambridge, MA 02139, USA. Correspondence should be addressed to G.S. (gregstep@mit.edu).

Published online 6 October 2004; doi:10.1038/nbt1016

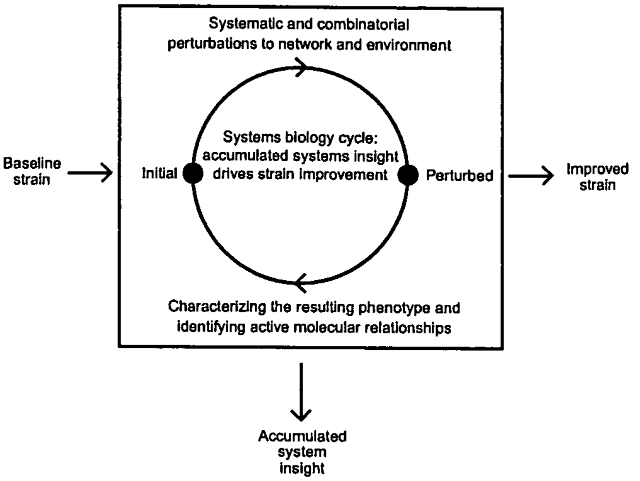

Systematic and combinatorial perturbations to network and environment
Systems biology cycle: accumulated systems insight drives strain improvement
Baseline strain
Initial
Perturbed
Improved strain
Characterizing the resulting phenotype and identifying active molecular relationships
Accumulated system insight

Figure 1 Accumulated system-insight drives the systems biology cycle for strain improvement. Iterative perturbations and systematic phenotype characterizations yield system insight through the integration of large data sets. A trade-off exists: more detailed characterization yields higher systems insight, but limits the number of perturbations that can be realistically evaluated.

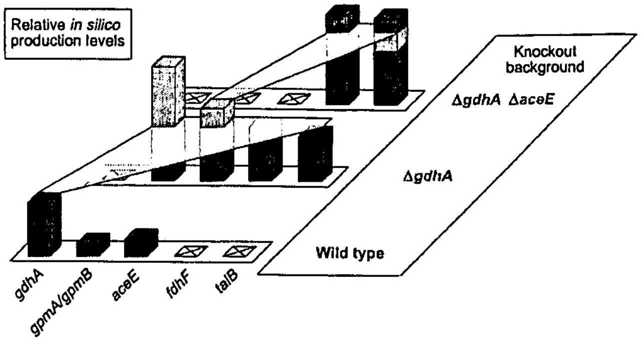


Figure 2 Identification of gene targets using global, stoichiometric modeling (H.A., Yong-Su Jin, J.M., G.S., unpublished data). This computational search makes use of a stoichiometrically balanced, genome-wide bioreaction network of *E. coli* metabolism whose fluxes are computed to maximize cell growth yield in the framework of Flux Balance Analysis and adjusted by MOMA. These results indicate that novel gene targets arise as the genotype is altered as a result of gene knockouts. This is especially evident in the case of *talB*. Although *talB* increases the production level in a *gdhA/aceE* knockout background, it is detrimental in a *gdhA*-only knockout background. The gene *ghdA* encodes glutamate dehydrogenase, *gpmB* encodes phosphoglucomutase, *aceE* encodes pyruvate dehydrogenase, *fdhF* encodes pyruvate formate lyase and *talB* encodes transaldolase.

Second, characterizing of the resulting phenotype allows identification of specific, differentially active molecular relationships. Using high-throughput methods, we may compare the initial and perturbed biomolecular landscapes on the gene, protein and small-molecule level. Such comparisons can reveal hundreds or thousands of molecules that vary significantly. High-level integration of biomolecular networks and states may identify concrete molecular relationships active under genetic and environmental perturbations. By specific, low-level exploration of the active biomolecular relationships, new perturbations to introduce can be identified[3].

Advances in high-throughput assays greatly enhance our ability to characterize the cellular phenotype. Sophisticated methods exist not only for analyzing and quantifying metabolite profiles, proteins and gene expression, but also for cataloging the interactions among network components. Applying these tools to a microbial system provides a detailed snapshot of cellular function.

However, one quickly encounters a trade-off between increased system characterization and the number of perturbations that can be realistically studied. Typically, the iterative strain-improvement process turns over much more rapidly than a fundamental systems biology investigation. At one extreme, one sees directed evolution in which randomly mutated strains with desired properties are preferentially selected and further modified with minimal evaluation. Here, the cellular characterization and system-wide understanding comprise only a single measurement of desired phenotype. On the other hand, even for those industrial organisms with sequenced genomes, the direct return on investment for any detailed and broad state characterization, such as gene expression profiling, can be difficult to ascertain.

Strain improvement and systems biology

Systems biology is a useful adjunct to traditional industrial programs aiming to design microbes optimized for maximal product formation. In addition to generating robust production strains, these applications often involve an important component of reverse engineering, whereby microbes with attractive properties are dissected for the purpose of transferring insights learned from their functions to the further improvement and optimization of production strains.

Global models for identification of gene targets. Advances in molecular biology for introducing genetic modifications at the molecular level have not been matched by equally effective approaches for identifying specific gene targets whose modification would bring about a desired phenotype, such as product overproduction. As a result, most strain improvement programs have resorted primarily to *ad hoc* or random approaches, whereby genetic perturbations are introduced and their effect on the phenotype of interest is evaluated along with other aspects of the resulting physiology.

The main obstacle to the rational solution of this problem is the lack of a reliable, global, metabolic model that captures the majority of the stoichiometric, kinetic and regulatory effects on metabolite interconversions and metabolic flux distributions through the cellular bioreaction network. Given the availability of genome sequences, the model closest to the above ideal is a global network of metabolic reactions described by detailed flux balances for each metabolite pool. This is usually an underdetermined system, and an *a priori* calculation of metabolic fluxes requires the optimization of a certain objective function, for example, growth yield maximization[4]. Although such 'maximum growth fluxes' do not necessarily represent the state of metabolism, they nevertheless provide insights into the distribution of carbon and energy resources required for optimal growth. A variation of the above approach, minimization of metabolic adjustment[5] (MOMA), attempts to determine more realistic flux distributions from genomic data alone by calculating profiles that are intermediate between the wild-type optimal and gene-knockout-mutant optimal.

The above framework can be used to guide the choice of gene knockout targets. In one such application of the flux balance method (modified by MOMA), we scanned the entire *Escherichia coli* genome for single- and multiple-gene knockouts that would increase the yield of lycopene (H.A., Yong-Su Jin, J.M., G.S., unpublished data). Figure 2 illustrates the selection strategy of targets for single, double and triple knockout mutants following a sequential approach. This

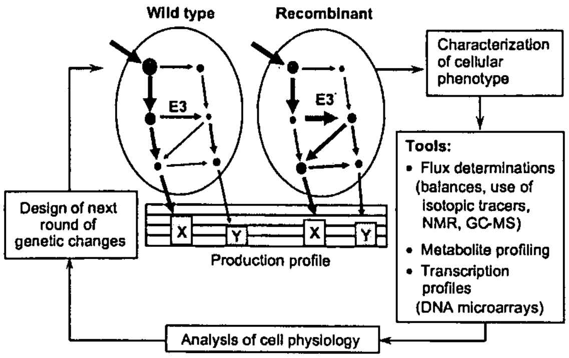
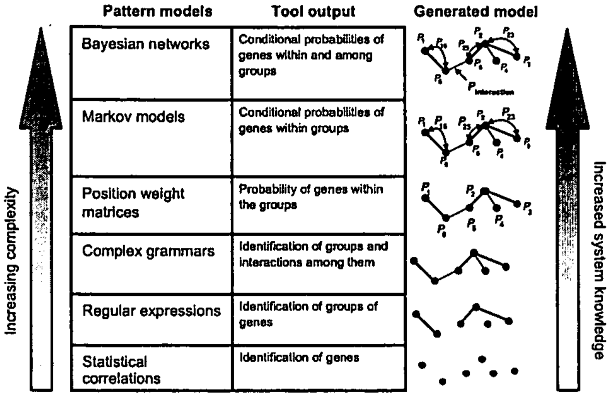

Wild type
Recombinant
E3
E3'
X
Y
X
Y
Production profile
Characterization of cellular phenotype
Tools:
• Flux determinations (balances, use of isotopic tracers, NMR, GC-MS)
• Metabolite profiling
• Transcription profiles (DNA microarrays)
Design of next round of genetic changes
Analysis of cell physiology

Figure 3 Exploiting complexities in metabolic networks. Metabolic engineering attempts to exploit the complexities of metabolic networks to improve cellular properties. In this schematic, wild-type cells are engineered to overexpress enzyme E3 with the goal of increasing the low yield of product Y. However, because of network interactions, overexpression of E3 has a minimal effect on the accumulation rates of either products Y or X. To improve the yield of product Y, multiple steps in the network will have to be targeted and genetically modified. In performing these steps, we gain insight into the biological system. Purple circles indicate the pool size of metabolites in the network. Arrow thickness depicts relative flux magnitude of the corresponding reactions.



Pattern models	Tool output
Bayesian networks	Conditional probabilities of genes within and among groups
Markov models	Conditional probabilities of genes within groups
Position weight matrices	Probability of genes within the groups
Complex grammars	Identification of groups and interactions among them
Regular expressions	Identification of groups of genes
Statistical correlations	Identification of genes

Figure 4 Approaches for pattern modeling useful in association analysis. Various pattern models and data analysis techniques may be used for the rigorous linking of datasets. In the example presented in the text, statistical correlations were used to link microarray data with phenotype (product formation). Models of increased complexity require more data, but can yield a much higher level understanding of mechanisms and underlying interactions. Furthermore, these techniques may be used to cross-validate proposed interaction networks and generate quantitative metrics (such as probabilities) for influential genes and their interactions.

method yielded a triple knockout that produced ~40% more lycopene compared with an engineered overproducing *E. coli* strain.

The success of the above approach justifies some optimism regarding more rational approaches to strain improvement but also raises a host of interesting questions. First, can we expect all gene knockouts identified by the above methods to show increased productivity? The answer is clearly no. The above computational approach simply identifies mutants with an increased availability of precursors for product formation, but as a purely stoichiometric calculation, it cannot incorporate regulatory or kinetic effects. Second, will such sequential, steepest ascent–like methods always cover the entire landscape? We clearly do not know. In this particular case, all possible gene pairs were investigated in an exhaustive computational search, and no pair that had not been covered by the sequential approach was found. This is a non-generalizable result, nevertheless. These observations underscore the importance of the iterative scheme of Figure 1 in elucidating the complex landscape of cellular function.

Randomized genetic tools, such as overexpression libraries and transposon mutagenesis, also allow broad-range perturbations of cellular systems. When coupled with targets identified from global modeling, these methods can be used for effective strain improvement. Furthermore, characterizing these systems allows the dissection of critical subnetworks within the cell. By investigating how product formation correlates with these regulatory networks, putative molecular interactions may be inferred and tested in subsequent perturbations.

Bioreaction network analysis. Metabolic engineering aims to improve strains using modern genetic tools. We showed that strains can be modified by introducing specific transport, conversion or deregulation changes that result in flux redistribution and product yield improvement[6]. Metabolic engineering differs from genetic engineering in that it is concerned with the entire metabolic system rather than with the overexpression of a gene. Although genetic engineering can successfully overexpress or otherwise modify a gene, this may have little impact on cell physiology. By examining the properties of the metabolic network in its entirety, metabolic engineering attempts to identify targets for amplification as well as to assess rationally the effect of such changes on the properties of the network. Metabolic engineering is thus a progenitor of functional genomics and systems biology in that it represents the first organized effort to reconstruct and modify pathways using genomic tools and guided by information about the behavior of the entire system[7]. This approach is illustrated by the concept of the distribution of kinetic control (metabolic control analysis, MCA)[6,8] and by the modification of multiple enzymes to achieve flux amplification, as we recently verified experimentally[9].

Figure 3 shows how the general paradigm of Figure 1 is applied to the specific case of metabolic engineering. A key point here is that the first rounds of genetic changes, though rarely successful, can nevertheless contribute invaluable insights that can guide the strain improvement cycle to a successful conclusion. Figure 3 also lists the tools available for more elaborate evaluation of strain physiology after a perturbation, although how these tools can be used systematically for strain improvement is still unclear. Nevertheless, we have shown that flux determination is an essential component of strain evaluation for metabolic engineering[10]. The examples below proceed from the modification of simple, linear product pathways to the perturbation of complex interacting networks encountered in central carbon metabolism.

We have investigated the production of threonine in a lysine-producing strain[11]. Carbon flux from aspartate is distributed between the lysine and the threonine-isoleucine pathways at the aspartate semialdehyde (ASA) branch point. There are three reactions from ASA to threonine, and the product of the first one, homoserine dehydrogenase (HDH), is feedback-inhibited by threonine. The identification of a mutant gene encoding a threonine-insensitive HDH solved the feedback-inhibition problem, but its overexpression succeeded in redirecting only 50% of the aspartate flux towards threonine. The reason was an imbalance between the activities of HDH and the next enzyme in the pathway, homoserine kinase (HK). When these two enzymes were balanced in a new construct that incorporated inducible control of the HK expression, almost 100% of the flux was redirected toward threonine synthesis. In this case, enzyme activity measurements along with enzymatic kinetic data about the sequential pathway reactions were the additional information that guided subsequent research.

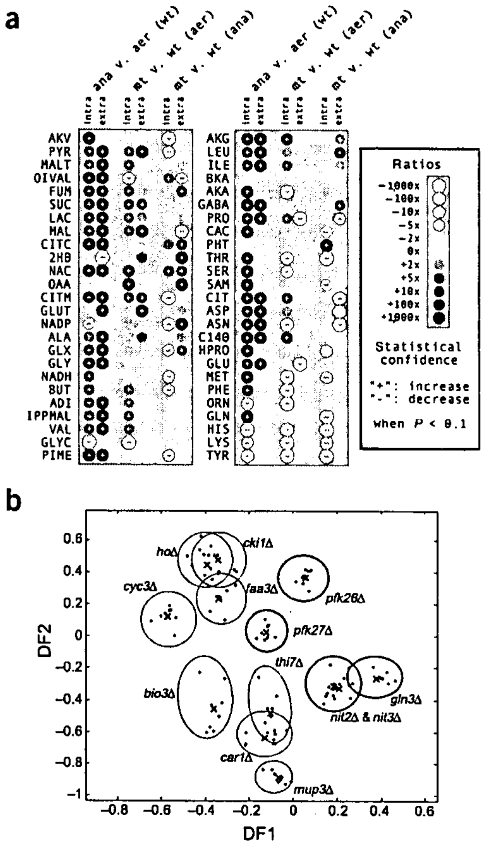


Figure 5 Use of metabolic state assays for phenotypic characterization. **(a)** GC-MS measurements of a broad diversity of intracellular metabolites 38 (J.M. and S. Villas-Boas, unpublished data). Here, we observe a ratio for each metabolite for three environmental and genetic perturbations (environmental: aerobic and aerobic cultivation, genetic: wild-type yeast and *gdh1* knockout), intracellularly and extracellulary. Statistically significant differences in metabolite pools are indicated by a '+' or '–' to indicate increases or decreases, respectively. By comparing these metabolite levels across samples, we may infer activated genetic and metabolic pathways. For instance, the anaerobic cultivations accumulate much higher levels of TCA cycle intermediates because the corresponding TCA cycle reactions occur at slower rates. **(b)** GCMS peaks for metabolite data may be clustered to classify mutants[37]. Here, 19 yeast deletion strains with various knockouts affecting central carbon metabolism and amino acid production were cultivated overnight in microtiter plates. From GC-MS data for the extracellular metabolites, a degrees of analysis (DFA) model was trained with 20 principal components. Based upon the first two principal components, the plot demonstrates clusters of samples that correspond to certain mutants In other words, the metabolic footprint of gas chromatography-mass spectrometry peaks contains enough phenotypic information to differentiate among mutants.

In a larger network application, modification of central carbon metabolism was considered for overproduction of aspartic-acid family amino acids, such as lysine. Single-gene overexpression had only marginally increased production[12], but flux control coefficient calculations revealed that most of the kinetic control for lysine overproduction resided in the lysine pathway[12]. Additionally, as shown by our laboratory, isotopic tracer probes applied to well-designed genetic backgrounds identified pyruvate carboxylase as the key reaction supplying more than 90% of the lysine carbon[13]. This led to the sequencing of pyruvate carboxylase in *Corynebacterium glutamicum*; however, its overexpression only marginally increased the yield of lysine. Further investigation identified aspartokinase as the bottleneck enzyme in a pyruvate carboxylase–overexpressing strain: when aspartokinase was simultaneously overexpressed with pyruvate carboxylase, the specific lysine productivity rose by one- to threefold depending on the carbon source used. In this example, detailed flux calculations provided the critical information that was best used in an MCA-defined framework.

The above case studies represent the different types of input information sought to explain the observed strain physiology and to rationally guide strain improvement. There are numerous other applications of this rapidly expanding field, whose very definition is linked with systems biology and which has produced impressive results in the past decade[14–29].

Integrating physiological and transcriptional data. The rapid adoption of new technologies suggests that future developmental programs will have the benefit of large volumes of data for the characterization of extensive libraries of strains. A key question is how to use these data most profitably to identify gene targets for modification. A related issue of critical importance is how to link data from the expression and metabolic phenotypes to help reveal the genetic basis of strain performance and organism physiology in general.

Various mathematical constructs are available to create links between large data sets and phenotypes. In one such application, association discovery was used to evaluate a library of unsequenced fungal strains of *Aspergillus terreus* for their ability to overproduce the antibiotic lovastatin[30]. First, using gene overexpression, a large number of strains with diverse lovastatin and (+)-Geodin production profiles was generated, and the strains were characterized by metabolite and transcriptional profiling. From this wealth of biological data, Askenazi *et al.*[30] extracted key putative parameters and genes by a statistical association analysis. To do this, each transcriptional data ratio and metabolic profile ratio (both calculated using the parental strain as basis) was correlated with the product level using Pearson correlations and principal components analysis (PCA). These statistical tools, applied to the sign of the correlation coefficient, provided a measure of the relative strength and impact (either positive or negative) that each transcript had on the product level. The analysis revealed key trends for some genes associated either positively or negatively with the production of lovastatin and (+)-Geodin. This led to the identification of genes whose modulation alters lovastatin production and eventually to an improvement in production of over 50%.

Techniques for linking data sets are not limited to simple statistical metrics, as illustrated in the above example. Various pattern discovery and characterization tools can yield different outcomes and levels of knowledge about the systems. **Figure 4** summarizes some commonly used pattern models and the typical system detail extracted from such analyses. Given enough data, one can extract probabilistic models that completely capture cellular interactions. In this way, relevant biomolecular networks may be discovered and manipulated without prior knowledge of an interaction network, which is especially important for unsequenced or poorly studied microorganisms. These concepts are illustrated in Box 1.

Use of metabolites for phenotype characterization. Strain phenotype characterization has relied primarily on transcript abundance and protein measurements. Only rarely have small metabolites been included in the measurement set, reflecting the difficulties in sampling and analyzing these molecules owing to their rapid time scales of change, especially at the high-throughput rates envisioned by the systems biology cycle of Figure 1. However, it is now accepted that, by

Box 1 Identification of active biomolecular pathways

Researchers continue to expand and refine pathway databases, but how does one determine which of the molecular relationships that define a mechanism might be relevant under a particular set of conditions? In one sense, pathway databases reflect a road map of possible functional routes connecting cellular components— but it is a static, noisy, incomplete road map that by itself yields limited understanding of how a perturbation induces a particular cellular phenotype. Clearly, we want to dissect how pathways are activated or repressed when a perturbation is introduced to identify targets for genetic manipulation.

Entries in pathway databases vary considerably in quality and reflect observations from either specific, bottom-up studies or global, top-down interaction screens. For instance, years of studying the galactose genetic switch has yielded a wealth of literature and very accurate interaction data about core regulation. Other connections, however, might be less well validated. *E. coli* pathway databases are based exclusively on knowledge accumulated from years of bottom-up studies. For yeast, however, the knowledge from bottom-up studies is supplemented by global experimental data. For protein-protein binding data, yeast two-hybrid[31] and protein-complex mass spectrometry experimentation have provided multifold coverage of each pairwise combination of protein interactions. Likewise, for transcription factor–mediated gene regulation, chromatin immunoprecipitation[32] microarrays measure DNA promoter binding sites genome-wide. These top-down approaches reveal interactions—albeit for a limited collection of organisms because of high costs—that have not been the focus of a specific, bottom-up study.

The determination of active biomolecular interactions can be facilitated with the Cytoscape software platform, which allows the integration of transcriptional data through global, biomolecular-interaction databases to identify the most active pathways[33]. Ideker *et al.*[34] demonstrated that differential (transcriptional) state measurements placed in the context of biomolecular networks can generate valuable pathway insight. Using galactose-pathway perturbations, they showed that putatively interacting genes were more likely to be differentially active together. This result reinforces the concept that pathways are activated and deactivated in a coordinated fashion. The Cytoscape platform allows visualization of these differential state data on the biomolecular network. Often, the user inputs differential gene expression data and uses the curated Biomolecular Interaction Network Database (BIND) for protein-protein (usually signaling) and protein-DNA (usually regulatory) interactions. Going a step further, the ActiveModules plug-in searches the biomolecular networks to determine which pathways are most affected by a set of perturbations[35]. By processing global data, one finds potentially interesting pathways that may be further explored through lower-level, detailed modeling and experiments.

Thus, a Cytoscape analysis facilitates the completion of the systems biology cycle by allowing the identification of active molecular relationships to target with the next iteration's set of perturbations. **Figure 6** illustrates a specific example in which differential transcription data reveal the pathways activated and deactivated upon the removal of the GAL80 repressor protein.

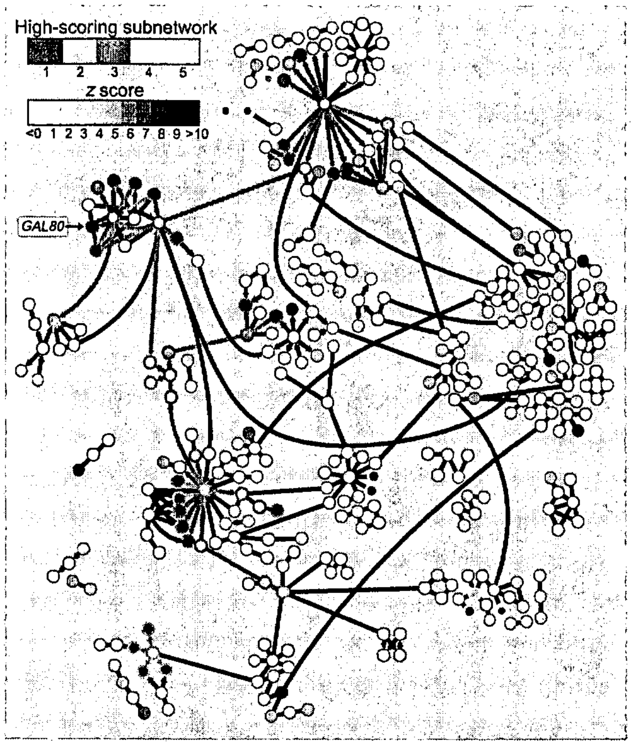


Figure 6 Determining active pathways after removing a transcription factor repressor. Identification of active pathways helps define gene targets. In this experiment, the GAL4 repressor gene GAL80 was deleted. Using microarray data superimposed on a predetermined set of protein-protein (pp) and protein-DNA (pd) interactions for yeast, the differential gene expression after gene deletion reveals the corresponding activated subnetwork illustrated above. Even without galactose present, removal of the GAL80 triggers cellular galactose-processing pathways by eliminating the repression of the GAL4 transcription factor. Node color indicates differential expression statistical significance for the particular gene, whereas node outline color and interaction edges between nodes indicates activated subnetworks. Significance of differential expression does not distinguish between upregulation and downregulation states; thus, both GAL80 (here, eliminated) and GAL1 (here, upregulated after GAL80 removal) will possess high z-scores as differentially active. As is evident from this figure, a single modulation of a gene can have a cascade effect throughout the biomolecular interaction network.

The ActiveModules plug-in scores each gene node and then searches, scores and compares possible active subnetworks. We note that, although the primary active pathway includes *GAL80* and the surrounding genes, not all active pathways directly connect to *GAL80*. This observation highlights the noisy, incomplete nature of the biomolecular networks. Often, one cannot directly trace the cascading effects of a perturbation because only a subset of biomolecules (e.g., genes but not small molecules) and their interactions have been considered. However, by identifying relationships that become active after certain perturbations, one gains valuable insight regarding rational gene targets to disrupt at the next iteration.

globally assaying metabolic states, we can identify a more diverse set of active molecular relationships indicative of mechanisms that encompass stoichiometry *and regulation*, as small metabolites are critical in regulating higher-level (transcriptional, translational) processes. Unfortunately, small molecules possess a wider range of chemical characteristics than do transcripts and are more difficult to measure simultaneously.

Gas chromatography coupled to mass spectrometry allows high-throughput analysis at relatively low cost. The gas chromatograph separates metabolites, whereas the mass spectrometer identifies and quantifies the metabolites corresponding to a given peak. Metabolite profiles in the culture supernatant are used to define a footprint of metabolic processes that occurred intracellularly. Such metabolite profiling will be applied increasingly to better define the cellular phenotype.

How will the observed metabolic fingerprints or footprints profiles help elucidate the physiological state of organisms following some perturbation? Figure 5 illustrates the use of dimensional reduction methods for visualizing different physiological states in a lower-dimensional space. Fisher discriminant analysis[36] projection of the metabolic data collected for strains subjected to environmental (aerobic and anaerobic cultivation) and genetic (*gdh1* knockout) perturbations succeeds in classifying the samples collected under different conditions and from different strains. Supervised and unsupervised classification methods such as these will be important in defining the location of a desirable metabolic state (such as one of high productivity) and in developing bioreactor controls that lead the system to the desired state in the course of a process.

Metabolic data can also be incorporated into databases that integrate transcription, protein-protein or protein-DNA interactions, and metabolism to identify biomolecular subnetworks that become activated in response to a perturbation. This vision entails an expanded Cytoscape framework encompassing metabolism along with transcriptional and higher-level processes in the cellular hierarchy. This broader metabolic state characterization will allow better understanding of the interplay between different pathways and will enhance the confidence of mechanism identification through the use of an expanded and more diverse set of measurements. As such, these methods have a role in the reverse engineering of strains.

Words of caution

The plethora of new data and new analytic methods justifiably leads to ambitious goals for systems biology. After many years of focused reductionist research, a deluge of information is expanding the scope of most investigations and promises to transform glimpses into snapshots of a dynamic world where cells grow, divide and produce, or organisms develop, differentiate and begin to deviate from the norm. No one can deny the opportunities that present themselves, but one must also be mindful that the problem that we set out to address is several orders of magnitude larger than those with which we are familiar. Consequently, it is important that we temper our expectations of immediate results and not lose sight of the following points:

First, despite the wealth of available genomic data, we are still unaware of numerous genes involved in important interactions and processes. A common misconception is that the genomic effort and accompanying analysis are almost complete. However, as recent results have demonstrated, this could not be further from the truth: genomic maps are continuously updated by discarding or adding (occasionally substantial amounts of) genes, proteins and new biomolecular interactions for pathways that were considered well understood.

Second, constructing biomolecular networks demands significant resources and expertise. Biomolecular networks incorporate a multitude of relationships connecting several types of components. At the genome scale, interaction maps require large experimental investments and subsequent analysis and curation. For instance, global protein-protein interaction maps exist for only a handful of model species, and reconstructing well-studied and well-documented networks, such as metabolic pathways, in a genome context requires years of curation.

Third, completeness is an elusive goal even for the better understood biomolecular networks. As more genomes are sequenced, the effort to uncover the structure and function of genetic regulatory networks has given birth to many databases, each of which attempts to distill the most salient features from incomplete and at times flawed knowledge. We mentioned the scant agreement among individual yeast two-hybrid screens. Clearly, 'accepted' interactions vary significantly across databases and over time, even for yeast, the best-studied system. Furthermore, many databases do not distinguish among direct and indirect interactions, raising uncertainty about the design of experiments to disrupt such interactions.

Fourth, high-throughput analytical tools, such as DNA microarrays, are widely available only for conducting measurements at the transcriptional level, and even these require training and incur significant costs. Going further to measure protein levels, protein states, regulatory elements, metabolites and metabolic fluxes requires complex and specialized equipment and significant expertise. Consequently, partnerships and collaborations are a *sine qua non* for effective research in systems biology.

Finally, patience is advised, as the more complex hypotheses derived from systems approaches are disproportionately harder to validate. Typically, after a perturbation, differential gene expression and network searching reveal active putative biomolecular networks that span dozens, if not hundreds, of genes. Figure 6 shows that even for a small network, a single gene removal causes a cascade of activation and deactivation among many genes. Although we can demonstrate cause and eventual effect, verifying the specific mechanism of each gene's activation/deactivation becomes a large task.

Looking forward

The reorientation of the research frame of mind to accept systems approaches to biological research will lead to better technologies for probing cellular phenotypes at ever increasing levels of accuracy and resolution. One can safely assume that the present trend of high-throughput measurement development will continue unabated, addressing molecules that are more difficult to measure and incorporating exciting new micro- and nanotechnologies.

More and more measurements, however, will not ensure faster progress toward the above objectives in the absence of effective tools for increasing the information content of measurements. Without underestimating the challenges in the development of measurement technologies, we believe that the key limiting step will be the development of mathematical-computational approaches that organize and integrate data to answer specific biological and biotechnological questions. This article has reviewed methods for meeting this fundamental challenge. It is our hope that it will help engage systems-oriented researchers and thus spur more activity in this area.

ACKNOWLEDGMENTS
We acknowledge the DuPont-MIT Alliance for research support and Kyle Jensen for his helpful suggestions.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Hemmi, H. *et al.* Identification of genes affecting lycopene formation in *Escherichia coli* transformed with carotenoid biosynthetic genes: candidates for early genes in isoprenoid biosynthesis. *J. Biochem.* **123**, 1088–1096 (1998).
2. Badarinarayana, V. *et al.* Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol.* **19**, 1060–1065 (2001).
3. Ideker, T. & Lauffenburger, D. Building with a scaffold: emerging strategies for high to low-level cellular modeling. *Trends Biotechnol.* **21**, 255–262 (2003).
4. Edwards, J.S. & Palsson, B. The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA* **97**, 5528–5533 (2000).
5. Segre, D., Vitkup, D. & Church, G.M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112–15117 (2002).
6. Stephanopoulos, G., Aristidou, A. & Nielsen, J. *Metabolic Engineering: Principles and Methodologies* (Academic Press, San Diego, 1998).
7. Stephanopoulos, G. & Vallino, J.J. Network rigidity and metabolic engineering in metabolite overproduction. *Science* **252**, 1675–1681 (1991).
8. Kacser, H. & Burns, J.A. The control of flux. *Symp. Soc. Exp. Biol.* **27**, 65–104 (1973).
9. Koffas, M.A., Jung, G.Y. & Stephanopoulos, G. Engineering metabolism and product formation in *Corynebacterium glutamicum* by coordinated gene overexpression. *Metab. Eng.* **5**, 32–41 (2003).
10. Stephanopoulos, G. Metabolic Fluxes and Metabolic Engineering. *Metab. Eng.* **1**, 1–10 (1999).
11. Colon, G.E., Nguyen, T.T., Jetten, M.S., Sinskey, A.J. & Stephanopoulos, G. Production of isoleucine by overexpression of *ilvA* in a *Corynebacterium lactofermentum* threonine producer. *Appl. Microbiol. Biotechnol.* **43**, 482–488 (1995).
12. Cremer, J. *et al.* Control of the lysine biosynthesis sequence in *corynebacterium glutamicum* as analyzed by overexpression of the individual corresponding genes. *Appl. Environ. Microbiol.* **57**, 1746–1752 (1991).
13. Koffas, M.A., Jung, G.Y., Aon, J.C. & Stephanopoulos, G. Effect of pyruvate carboxylase overexpression on the physiology of *Corynebacterium glutamicum. Appl. Environ. Microbiol.* **68**, 5422–5428 (2002).
14. Berrios-Rivera, S.J., Bennett, G.N. & San, K.-Y. The effect of increasing NADH availability on the redistribution of metabolic fluxes in *Escherichia coli* chemostat cultures. *Metab. Eng.* **4**, 230–237 (2002).
15. Cameron, D.C. *et al.* Metabolic engineering of propanediol pathways. *Biotechnol. Prog.* **14**, 116–125 (1998).
16. Causey, T.B., Shanmugam, K.T., Yomano, L.P. & Ingram, L.O. Engineering *Escherichia coli* for efficient conversion of glucose to pyruvate. *Proc. Natl. Acad. Sci. USA* **101**, 2235–2240 (2003).
17. Farmer, W.R. & Liao, J.C. Improving lycopene production in *Escherichia coli* by engineering metabolic control. *Nat. Biotechnol.* **18**, 533–537 (2000).
18. Fussenegger, M. & Betenbaugh, M.J. Metabolic engineering II. Eukaryotic systems. *Biotechnol. Bioeng.* **79**, 509–531 (2002).
19. Kaup, B. *et al.* Metabolic engineering of *Escherichia coli*: construction of an efficient biocatalyst for D-mannitol formation in a whole-cell biotransformation. *Appl. Microbiol. Biotechnol.* **64**, 333–339 (2004).
20. Khosla, C. & Bailey, J.E. Heterologous expression of a bacterial hemoglobin improves the growth properties of recombinant *Escherichia coli. Nature* **331**, 633–635 (1988).
21. Lee, S.Y. & Lee, Y. Metabolic engineering of *Escherichia coli* for production of enantiomerically pure (*R*)-(-)-hydroxycarboxylic acids. *Appl. Environ. Microbiol.* **69**, 3421–3426 (2003).
22. Martin, V.J.J. *et al.* Engineering the mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.* **21**, 796–802 (2003).
23. Nakamura, C. & Whited, G. Metabolic engineering for the microbial production of 1,3-propanediol. *Curr. Opin. Biotechnol.* **14**, 454–459 (2003).
24. Ostergaard, S. *et al.* Increasing galactose consumption by *Saccharomyces cerevisiae* through metabolic engineering of the GAL gene regulatory network. *Nat. Biotechnol.* **18**, 1283–1286 (2000).
25. Snell, K. & Peoples, O. Polyhydroxyalkanoate polymers and their production in transgenic plants. *Metab. Eng.* **4**, 29–40 (2002).
26. Stafford, D.E. *et al.* Optimizing bioconversion pathways through systems analysis and metabolic engineering. *Proc. Natl. Acad. Sci. USA* **99**, 1801–1806 (2002).
27. Stephanopoulos, G. & Kelleher, J. Biochemistry: how to make a superior cell. *Science* **292**, 2024–2025 (2001).
28. Viswanathan, K. *et al.* Engineering sialic acid synthetic ability into insect cells: identifying metabolic bottlenecks and devising strategies to overcome them. *Biochemistry* **42**, 15215–15225 (2003).
29. Xue, Y. & Sherman, D.H. Alternative modular polyketide synthase expression controls macrolactone structure. *Nature* **403**, 571–575 (2000).
30. Askenazi, M. *et al.* Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol.* **21**, 150–156 (2003).
31. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
32. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science* **298**, 799–804 (2002).
33. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
34. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
35. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics* **18** (suppl. 1) 233–240 (2002).
36. Stephanopoulos, G., Hwang, D., Schmitt, W.A., Misra, J. & Stephanopoulos, G. Mapping physiological states from microarray expression measurements. *Bioinformatics* **18**, 1054–1063 (2002).
37. Allen, J. *et al.* High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.* **21**, 692–696 (2003).